Generalized Composite Motif Discovery

Geir Kjetil Sandve and Finn Drabløs

Norwegian University of Science and Technology, 7052 Trondheim, Norway
{Sandve,Finn.Drablos}@ntnu.no

Abstract. This paper discusses a general algorithm for the discovery of motif combinations. From a large number of input motifs, discovered by any single motif discovery tool, our algorithm discovers sets of motifs that occur together in sequences from a positive data set. Generality is achieved by working on occurrence sets of the motifs. The output of the algorithm is a Pareto front of composite motifs with respect to both support and significance. We have used our method to discover composite motifs for the AlkB family of homologues. Some of the returned motifs confirm previously known conserved patterns, while other sets of strongly conserved patterns may characterize subfamilies of AlkB.

1 Introduction

Motif discovery in DNA and protein sequences is an important field in bioinformatics. Unique motifs found in a set of related sequences are often associated with the biological activity of the sequences. Motifs representing active site residues in enzymes (proteins) or transcription factor binding sites in genomes (DNA) are typical examples. Such motifs can also be used for classification of novel sequences or sequences outside the original training set. Both probabilistic and deterministic approaches are used. Arguably, deterministic approaches give the most easily interpretable results, as they represent motifs e.g. by subsets of regular expressions that either match a given sequence or not.

There are many different algorithms for motif discovery, including manual approaches. The earliest algorithms had very limited expressibility and could only discover substrings of amino acid symbols. PROSITE[1], a database of manually annotated motifs, in many ways set the standard for expressibility of deterministic motifs for proteins. In addition to exact symbols, PROSITE patterns also consist of fixed gaps, flexible gaps and ambigous symbols. Most automated motif discovery tools are only able to discover motifs consisting of a subset of these components.

The discovery of motif combinations is an area of active research, for which both probabilistic and combinatorial approaches are used. Gibbs sampler[2] and PRINTS[3] are two well-known probabilistic approaches. Most combinatorial approaches discover spaced dyads[4][5] or ordered sets of motifs with strong distance constraints[6]. Brazma et al.[7] are among the few methods that discover unordered sets of motifs.

A set of single motifs is a general starting point for composite motif discovery. Many advanced methods exist for the discovery of single motifs, and none are

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 763–769, 2005.

superior in all respects[8]. We have therefore chosen to develop an algorithm for the discovery of motif combinations that can use single motifs generated by any deterministic motif discovery tool.

GCMD (Generalized Composite Motif Discovery)[1], exhaustively identifies the most significant combinations of a set of precomputed motifs. It can be set to discover both ordered and unordered motifs, with or without distance constraints. In addition to being flexible with regards to both single and composite motif model, and exhaustive in search for combinations, two properties clearly distinguish our algorithm from previous approaches: we model the problem as a two-goal optimization with the optimal Pareto front as output, and we automatically discover potential subfamilies.

GCMD is here discussed mainly in terms of protein sequence motifs. However, the tool itself is general and can also be applied to motifs from DNA sequences.

2 The Generalized Composite Motif Discovery Tool

In broad terms, GCMD takes as input a set of single motifs and exhaustively discovers the optimal motif combinations with respect to both support and significance. This is more thoroughly explained in the following sections.

2.1 Vocabulary

The set of sequences that have at least one occurrence of a given motif, is called the *occurrence set* of the motif. The cardinality of the occurrence set is referred to as *support*.

We use the term *single motifs* to denote the motifs that are input to the GCMD algorithm, and *composite motifs* to denote the discovered motifs that are sets of single motifs. The term *component* is used to denote one of the single motifs that makes up a composite motif.

We also use the terms *Pareto domination* and *Pareto front* in multiple criteria optimization. A motif is Pareto dominated if there exists another motif having equal or higher values of both support and significance, where one of the values is strictly higher. Since support is a discrete value, this means that a motif is Pareto dominated if there exist another motif with equal or higher support, and strictly higher significance. The Pareto front is the set of all non-dominated motifs. In our case this is the most significant motif for each value of support.

2.2 Motif Representations

The first step in using GCMD is to discover deterministic single motifs with a separate motif discovery tool. For tools that discover probabilistic motifs, a threshold may be used to make them deterministic. A bitstring is then constructed for each motif, where the i'th bit is 1 if the motif has an occurrence in

[1] The code is available upon request to first author

the i'th sequence, and 0 otherwise[7]. A composite motif occurs in a sequence if, and only if, every single motif in the set occurs in the sequence. This leads to a basic representation of a composite motif as a set of indexes to its component motifs, as well as an occurrence set calculated by taking the intersection of the occurrence sets of all component motifs.

2.3 Significance Evaluation

Significance of motifs is measured as negative log-likelihoods, using the same calculations as the motif discovery method Splash[9]. More specifically, the significance of a single motif is the negative log-likelihood of observing the motif in a random background sequence with the same amino acid distribution as the input sequences. As single motifs usually are short compared to sequence length, the log-likelihood of a composite motif is in general well approximated as the sum of log-likelihoods of its components.

2.4 Significance vs Support

Both significance as well as support is important when evaluating motifs, and it is not easy to make the right trade-off between these properties when doing automated motif discovery. Most algorithms require a threshold on support, and this threshold is often user specified. Using a very strict value may lead to loss of significant motifs that are characteristic of subfamilies of sequences. On the other hand, a too permissive threshold may lead to searches dominated by motifs with high statistical significance in subsets of sequences, and one may lose less significant motifs representing weak commonalities characteristic of larger sequence families.

By formulating the motif discovery problem as a two goal optimization, we can explore a very large search space of interesting motifs, and return information about this in a condensed form as a Pareto front. The user gets a diverse set of motifs, and can readily see the tradeoff between significance and support as the number of sequences taken into consideration increases. This removes the need to set explisit thresholds on support or significance.

2.5 Pruning of Search Space

GCMD traverses the search space exhaustively and returns the set of Pareto optimal composite motifs. The size of the search space is $\binom{n}{c}$, where n is the number of single motifs used as input to GCMD, and c is the desired number of components in the composite motifs. Many algorithms exist for the mining of frequent item sets. Brazma et al.[7] uses the algorithm of Toivonen[10] to discover unordered sets of motifs. As this algorithm do pruning only based on support, it can not handle the large number of input motifs and low values of support that we are interested in.

We have developed a branch-and-bound algorithm, tailored to our two goal optimization problem, that is very efficient on real biological data. Since our

goal is to find an optimal Pareto front with respect to support and significance, we need to determine upper bounds on both of these values. An upper bound on support is simply the minimum support of the current components of the composite motif. To introduce an upper bound on significance, we ensure that when a composite motif is expanded, the new component has a lower significance value than all other components of the motif. Note that this does not reduce the set of composite motifs we are able to discover, it only excludes all but one of the $n!$ permutations that corresponds to the same combination of n single motifs. For a given motif c_i, this leads to a straightforward upper significance bound on any expansions of c_i with n components :
$s(c_n) \leq s(c_i) + (n - i) * s(c_i(i))$, where c_i is a motif with i components, $c_i(i)$ is the i'th component of motif c_i, and $s(c)$ is the significance of motif c.

With these upper bounds in place we can make a recursive function that takes as parameter a composite motif c that is to be expanded. For each single motif s with significance lower than all current component significances of the motif, we check whether the resulting upper bounds on support and significance are dominated by the current Pareto front. If not, a new composite motif is formed from c, with the single motif s as an added component. The resulting motif is stored in the Pareto front if it has reached the desired number of components, otherwise it is again expanded recursively.

In order to reduce the number of explored composite motifs even further, we explore the expansions of a given composite motif in order of decreasing significance of single motifs. Note that the support of the composite motif before any new expansion is an upper bound on support. As the upper bounds are monotonically decreasing, we can stop exploring new expansions of a composite motif as soon as the upper bounds on support and significance are dominated by the Pareto front.

2.6 Automated Subfamily Discovery

The Pareto front of composite motifs for a family may contain significant motifs with relatively low values of support. It is natural to ask whether such a motif characterize a subfamily of the data set. One may therefore try to discover new motifs in the sequences that are not in the occurrence set of the first composite motif. Since the goal is to find motifs that are common to as many sequences as possible, we have restricted automated subfamily discovery to only two subfamilies and also demand that one of the motifs belong to the Pareto front of the whole family. Significance values of motifs are log-likelihoods, and a two-subfamily-motif occur in a given sequence if either of the one-subfamily-motifs occur in the sequence. Therefore, the significance of a 2-subfamily-motif c is well approximated as: $s(c) = log_2(2^{s(c_a)} + 2^{s(c_b)} + 2^{s(c_a)+s(c_b)})$, where c_a and c_b are the one-subfamily-motifs.

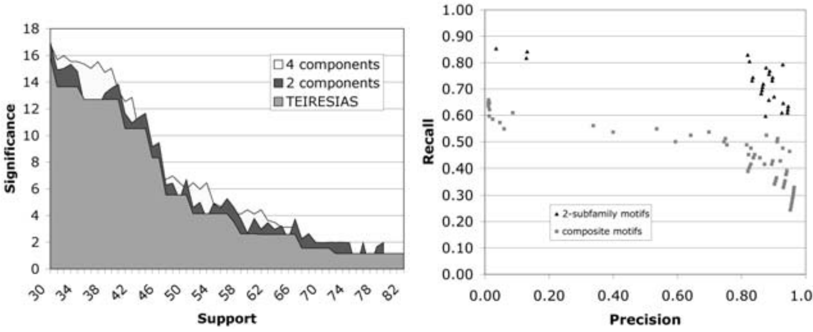
3 Results and Discussion

The family of AlkB homologues (ABHs) was used as a test case for composite motif discovery. The ABHs are members of the 2-oxoglutarate and Fe^{2+}-dependent

(2OG-Fe(II)) oxygenase superfamily[11]. They have been shown to be involved in repair of methylation damage of DNA and RNA through a direct reversal mechanism, where the methyl group is oxidised and spontaneously released as formaldehyde[12]. Recent screening of databases using sensitive search methods has shown that ABH-like sequences are widespread in bacteria and eukaryotes, see Drabløs et al.[13] for a review.

The degree of sequence conservation in the ABH family seems to be very low, basically just a `H.D` motif, an isolated `H` and a `R.....R` motif (using single-letter amino acid symbols) is completely conserved in most ABH alignments. All except the final `R` are involved in coordination of the Fe^{2+} ion, the final `R` is probably involved in substrate binding as it seems to be relatively unique to the ABH family of this superfamily[11]. However, there may be subfamilies within the ABH family with more extensive conservation, and there may be additional conserved patterns in sequence regions that are difficult to align correctly by traditional methods. The ABH family is therefore an interesting test case with practical implications.

A set of 82 AHB-like sequences, previously investigated in [13], was used for the analysis. Teiresias[14] was used to generate 50.000 single motifs from the input sequences, and GCMD was used to identify the Pareto front for composite motifs with 2 and 4 components, using chemical equivalence sets for residue types (Fig. 1(a)). The significance of the composite motifs is higher for most support values compared to single motifs. However, here GCMD is used mainly to identify interesting composite motifs and correlate this with biological significance. The dominating single motif, which is used in most of the composite motifs, is `[ILMV]..H.[DE]`. This corresponds to the first Fe^{2+} binding motif in the ABH sequences. In particular for composite motifs with high support this is mainly combined with variants of the motif `[KR]..[ILMV]..[KR]`, which corresponds to the Fe^{2+} and possibly substrate binding `R` groups. This shows that



(a) Pareto front of single and composite motifs for ABH. Significance is the negative $\log_2$-likelihood of a motif

(b) ROC of the discovered motifs for ABH. Recall is $TP/(TP + FN)$ and precision is $TP/(TP + FP)$)

Fig. 1.

the most interesting motifs identified by GCMD also have biological relevance, and that GCMD is able to identify such motifs from a large and complex set of input data.

However, it is evident that there are subfamilies of ABH-like sequences in the data set, and depending on the selected threshold for support several such subfamilies may be identified. One example is the composite motif `(L..G.[ILMV][ILMV].M....[QN]) & ([FY]....[DE].[ILMV]..H.D)`, which seems to be characteristic of the hABH2/hABH3 subfamily (human ABH type 2 and 3). This subfamily has been extensively studied experimentally[15]. As the detailed 3D structure of the ABH family still has not been experimentally determined, a detailed investigation of the biological relevance of these motifs probably has to be postponed until such data are available. However, this test shows that the GCMD method is able to identify biologically interesting subfamilies in a complex data set.

Although GCMD has not been developed as a classification tool, the classification performance may still serve to validate that the discovered motifs are indeed characteristic for a given family. Fig. 1(b) shows the receiver operating characteristic with respect to recall and precision when using the set of motifs in the Pareto front for classification. The introduction of subfamily motifs leads to a significant improvement in recall, and a larger fraction of the motifs have a high precision, compared to general composite motifs.

The performance of GCMD was also tested on 5 selected families from the PROSITE database. These PROSITE families are assumed to be difficult test cases, as the existing PROSITE patterns give low values for precision and recall. We used TEIRESIAS for single motif discovery. The Pareto front of composite motifs showed an average log-likelihood improvement of 20.4 compared to single motifs. The composite motifs in the Pareto front were used to classify the full set of SWISS-PROT[16] entries. For two of the five families (PS00485, PS00690) we were able to improve both precision and recall as compared to PROSITE, for two families we got comparable performance (PS00732, PS01048), and for the last family the PROSITE motif performed better (PS00187).

4 Conclusion

In our work we have built directly on previous work and focused on finding interesting combinations of single deterministic motifs discovered by separate motif discovery tools. Tests show that our tool is able to identify unique and biologically relevant composite motifs in very large data sets of single motifs.

Future directions of research include expanding the expressibility of deterministic motifs even further, as well as using the tool on other motif discovery problems, like for instance the discovery of transcription factor binding sites.

Acknowledgements

We want to thank Arne Halaas, Kai Trengereid, Magnus Lie Hetland, Rolv Seehuus, Tarjei Hveem and Øystein Lekang for help and suggestions. FD has been supported by FUGE/NFR (151899/150).

References

1. Bucher, P., Bairoch, A.: A generalized profile syntax for biomolecular sequence motifs and its fuction in automatic sequence interpretation. In: Proc Int Conf Intell Syst Mol Biol. 2 (1994) 53–61
2. Neuwald, A. F., Liu, J. S., Lawrence, C. E.: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. **4** (1995) 1618–1632
3. Attwood, T. K., Beck, M. E., Bleasby, A. J., Parry-Smith, D. J.: PRINTS - a database of protein motif fingerprints. Nucleic Acids Res. **22** (1994) 3590–3596
4. van Helden, J., Rios, A. F., Collado-Vides, J.: Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. **28** (2000) 1808–1818
5. Eskin, E., Pevzner, P.A.: Finding composite regulatory patterns in DNA sequences. Bioinformatics **18 Suppl 1** (2002) S354–S363
6. Marsan, L., Sagot, M.F.: Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. J Comput Biol. **7** (2000) 345–362
7. Brazma, A., Vilo, J., Ukkonen, E., Valtonen, K.: Data mining for regulatory elements in yeast genome. In: Proc Int Conf Intell Syst Mol Biol. 5 (1997) 65–74
8. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B.D et al.: Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23** (2005) 137–144
9. Hart, R.K., Royyuru, A.K, Stolovitzky, G., Califano, A.: Systematic and fully automated identification of protein sequence patterns. J Comput Biol. **7** (2000) 585–600
10. Toivonen, H.: Discovery of Frequent Patterns in Large Data Collections. PhD thesis, University of Helsinki (1996)
11. Aravind, L., Koonin, E.V.: The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. **2** (2001) RESEARCH0007
12. Falnes, P.O., Johansen, R.F., Seeberg, E.: AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli. Nature **419** (2002) 178–182
13. Drabløs, F., Feyzi, E., Aas, P.A., Vaagboe, C.B., Kavli, B., Bratlie, M.S., Peña-Diaz, J., Otterlei, M., Slupphaug, G., Krokan, H.E.: Alkylation damage in DNA and RNA–repair mechanisms and medical significance. DNA Repair **3** (2004) 1389–1407
14. Rigoutsos, I., Floratos, A.: Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics **14** (1998) 55–67
15. Aas, P., Otterlei, M., Falnes, P., Vaagboe, C., Skorpen, F., Akbari, M., Sundheim, O., Bjoras, M., Slupphaug, G., Seeberg, E., Krokan, H.: Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. Nature **421** (2003) 859–863
16. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, MC., Estreicher, A. et al.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31** (2003) 365–370.